A Hybrid Model for Co-reference Resolution

Jason D. M. Rennie
jrennie@csail.mit.edu

March 11, 2005

1 Introduction

We consider the problem of Co-reference Resolution. Given a text[1], we would like to be able to group together noun phrases (NP) that refer to the same entity. Past approaches can be separated into two groups. Early work (Aone & Bennett, 1995; Soon et al., 2001; Ng & Cardie, 2002) treated it as a classification problem—the goal being to determine the antecedent of each NP. More recently (Cardie & Wagstaff, 1999; McCallum & Wellner, 2003), it has been viewed as a clustering problem. In particular, McCallum and Wellner consider it as a clustering problem where the distance metric is unknown and must be learned. McCallum and Wellner correctly point out the need for consistency within clusters. Their example ("Mr. Powell" → "Powell" → "she") provides valuable motivation for their model. Their empirical results indicate that theirs is an excellent model for grouping proper NPs. The problem of grouping proper NP is well fit to their model since intra-cluster consistency is a primary constraint of the problem. Not so for non-proper NPs, where locality and sentence structure play larger roles. For non-proper NPs, the complete graph model with pairwise potentials is a poor model since long-range consistency is of negligible importance. Also, it may be a detriment to use the same set of weights for interactions between both proper and non-proper NPs. We believe that early work on classification provides a more compelling framework for non-proper NP reference resolution. They take the perspective of trying to identify the single antecedent for each NP. This is inappropriate for proper NPs, but well suited for non-proper NPs such as pronouns. We would like to integrate the classification approach with the clustering approach, jointly learning a distance function, but using a clustering-type objective on the proper NPs and a classification-type objective on the non-proper NPs. But, decision trees do not integrate well with probabilistic models. In place of decision trees, we insert a softmax gating network, where the set of experts is the NPs that have preceeded it in the text. i.e. the label distribution for a non-proper NP is a mixture over the label distributions of previous NPs, where the mixing weights are determined by softmax.

[1] We presume that the text is English; however, our approach is general enough to apply to a wide variety of languages.

2 The Model

Let x_i represent the i^{th} noun phrase (NP). Let A be the set of indices of proper NPs; let B be the set of indices of non-proper NPs (e.g. pronouns). Let y_i represent the entity to which x_i refers. We define a pairwise indicator variable,

$$y_{ij} = \begin{cases} +1 & \text{if } y_i = y_j \\ -1 & \text{if } y_i \neq y_j \end{cases} . \tag{1}$$

Our model involves two parts, a model on proper NPs, which is identical to that of McCallum and Wellner (2003), and a model on non-proper NPs. Given a set of indices, S, let y_S be the set of labels associated with those indices. We use Bayes' Law to write our model as the product of two models,

$$P(\vec{y}|\vec{x}) = P(y_A|\vec{x})P(y_B|y_A, \vec{x}). \tag{2}$$

For the proper NP model, we use Model 3 from McCallum and Wellner (2003). We define a similarity between pairs of NPs,

$$s(x_i, x_j) = \sum_k w_k f_k(x_i, x_j). \tag{3}$$

The feature functions, $\{f_k\}$, define the relation between NPs; the weights $\{w_k\}$ define how the features impact similarity. Training our model involves optimizing the set of weights. We also define pairwise potential functions, which are simply an exponentiated form of similarity that incorporates label agreement:

$$\psi(x_i, x_j, y_{ij}) = e^{y_{ij}\, s(x_i, x_j)}. \tag{4}$$

Then, the joint probability of the proper NP labels is simply a product of the potentials,

$$P(y_A|\vec{x}) = \frac{1}{Z_{\vec{x}}} \prod_{i,j \in A} \psi(x_i, x_j, y_{ij}), \tag{5}$$

where the normalization constant, $Z_{\vec{x}} = \sum_{y_A} \prod_{i,j \in A} \psi(x_i, x_j, y_{ij})$, sums over all possible configurations of labels for the proper nouns.

For the non-proper NP model, we make three assumptions: (1) each NP has a single antecedent, (2) an NP and its antecedent share the same label, and (3) the antecedent must come before the NP in the text. Let the labels for all NPs before the i^{th} (non-proper) NP be given. We define $B_i = \{j \in B | j < i\}$ to be the set of indices of non-proper NPs that come before x_i. Then the conditional probability for y_i is

$$P(y_i|y_A, y_{B_i}, \vec{x}) = \sum_{j<i|y_{ij}=1} \frac{\psi(x_i, x_j, 1)}{\sum_{j<i} \psi(x_i, x_j, 1)} \tag{6}$$

The product of non-proper NP conditionals gives us the second part of our model: $P(y_B|y_A, \vec{x}) = \prod_{i \in B} P(y_i|y_A, y_{B_i}, \vec{x})$.

The product of these two models gives us our joint model,

$$P(\vec{y}|\vec{x}) = \frac{\prod_{i,j \in A} \psi(x_i, x_j, y_{ij})}{\sum_{y_A} \prod_{i,j \in A} \psi(x_i, x_j, y_{ij})} \prod_{i \in B} \frac{\sum_{j<i|y_{ij}=1} \psi(x_i, x_j, 1)}{\sum_{j<i} \psi(x_i, x_j, 1)}. \quad (7)$$

3 Learning

Given a set of clustered data, we would like to learn weights to maximize the likelihood of the model. To simplify learning, we separate the features and weights according to the sub-models. We use one set of features/weights for the proper part of the model and a separate set of features/weights for the non-proper part of the model. What this means is that a proper NP feature is zero if at least one of the two NPs is non-proper; a non-proper NP feature is zero if both NPs are proper. By separating the weights/features in this way, we can optimize the two sub-models separately.

McCallum and Wellner (2003) have already discussed the issues in learning parameters for the proper NP part of the model. $P(y_A|\vec{x})$ is convex in the weights, but the normalization is a sum over exponentially many terms. Also, the second expectation in the gradient is a sum over exponentially many terms. We follow the lead of McCallum and Wellner and use stochastic gradient ascent in the form of a voted perceptron, a fast, but approximate method for calculating the gradient. The gradient for the non-proper sub-model is relatively simple to compute ($O(n^2)$), so we do not need to revert to approximate methods. However, due to the mixture nature of the model, it is non-convex. We use multiple starting points to partially alleviate the non-convexity.

4 Inference

For inference, we again assume a separation of the two sub-models. We use the approximate min-cut algorithms used by McCallum and Wellner to infer a labeling on the proper NPs. Given the proper NP labels, finding a maximum likelihood labeling of the non-proper NPs is straightforward and exact. If we were to not separate the models like this, more general-purpose algorithms, such as belief propagation would be necessary. Also, we would incur an undesirable bias towards fewer clusters; because of local normalization, the non-proper NP model prefers a single, global cluster.

References

Aone, C., & Bennett, S. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. *Proceedings of the 33rd conference on Association for Computational Linguistics*.

Cardie, C., & Wagstaff, K. (1999). Noun phrase coreference as clustering. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 82–89).

McCallum, A., & Wellner, B. (2003). Toward conditional models of identity uncertainty with application to proper noun coreference. *Proceedings of the IJCAI Workshop on Information Integration on the Web.*

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).*

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, *27*, 521–544.